DOCUMENT RESUME

ED 408 326 TM 026 579

AUTHOR Chiu, Chris W. T.; Wolfe, Edward W.
TITLE Generalizability Theory: A New Approach To Analyze Non-Crossed Performance Assessment Data.
SPONS AGENCY American Coll. Testing Program, Iowa City, Iowa.
PUB DATE Mar 97
NOTE 38p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Students; *Data Analysis; *Essay Tests; *Generalizability Theory; Higher Education; *Performance Based Assessment; Writing Tests
IDENTIFIERS *Missing Data

ABSTRACT

Unstable, and potentially invalid, variance component estimates may result from using only a limited portion of available data from operational performance assessments. However, missing observations are common in these settings because of the nature of the assessment design. This paper describes a procedure for overcoming the computational and technological limitations in analyzing data with missing observations by extracting data from a sparsely filled data set into analyzable smaller subsets of data. This parsing is accomplished by creating data sets that exhibit structural designs that are common in generalizability analyses, namely the crossed, mixed, and nested designs. An example of how to perform the procedure is given. Data are from a large-scale college writing assessment in which each of 5,905 examinees responded to 2 essay prompts. Results show that the sparsely filled performance assessment data sets can be restructured into analyzable smaller subsets of data. Results suggest that the crossed, mixed, and nested methods are comparable, but more study is needed to determine whether the methods generalize to other data sets with more than two facets. (Contains 3 figures, 9 tables, and 17 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Running Head: Analyzing Non-Crossed Performance Assessment Data

ED 408 326

Generalizability Theory: A New Approach to Analyze Non-Crossed Performance Assessment Data

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Chris W.T. Chiu

Michigan State University

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Chris W.T. Chiu
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Edward W. Wolfe

American College Testing, Iowa City, Iowa

Author Notes

Chris W.T. Chiu, Measurement and Quantitative Methods; Edward W. Wolfe (now at the Center for Performance Assessment, Educational Testing Service, Princeton, New Jersey), Performance Assessment Center.

Portions of this research were supported by the Summer Intern program at American College Testing. This manuscript was presented at the Annual Meeting of the American Educational Research Association, March, 1997 in Chicago, Illinois.

Correspondence concerning this article should be addressed to Chris W. T. Chiu, 163 Rampart Way Apt 202, E. Lansing, MI 48823. Electronic mail may be sent via Internet to chiuwing@pilot.msu.edu.

TM026579

BEST COPY AVAILABLE

Acknowledgments

The authors thank Robert Brennan and Dean Colton for the conceptualizations of some of the early analyses used in this paper. The authors also thank Bradley Hanson for his contribution on some of the programming, and Randall Fotiu for his consultation on estimation methods used in the analyses.

Finally, the authors are grateful to Betsy Becker, Yuk Fai Cheong, Robert Floden, Wen-Ling Yang, and members in the SynRG[1] research group at Michigan State University for their enlightening suggestions and constructive comments.

[1] The Synthesis Research Group (or SynRG) is a group of faculty and current and former graduate students who are interested in the development and application of quantitative methods for the synthesis of research results, often called meta-analysis. SynRG is also open to a variety of research topics.

Abstract

Unstable, and potentially invalid, variance component estimates may result from using only a limited portion of available data from operational performance assessments. However, missing observations are common in these settings because of the nature of the assessment design. This paper describes a procedure for overcoming the computational and technological limitations in analyzing data with missing observations by extracting data from a sparsely-filled data set into analyzable smaller subsets of data. This parsing is accomplished by creating data sets that exhibit structural designs that are common in generalizability analyses, namely the crossed, mixed, and nested designs. An example of how to perform the procedure is given.

Generalizability Theory: A New Approach to Analyze Non-Crossed Performance Assessment Data

Introduction

In recent years, performance assessment has become popular as a means for assessing students because these assessments provide direct measures of non-traditional student outcomes. Generalizability theory (G-theory), developed by Cronbach, Gleser, and Rajaratnam (1963), is often used in the development of performance assessments to identify the relative strengths of multiple sources of measurement error and to make projections concerning how to increase score reliability. A common problem encountered by those using G-theory with large-scale performance assessments is working with missing data (i.e., observations are missing for some pairings of the elements of two or more facets). The purpose of this paper is to investigate the comparability of several methods for analyzing data sets with missing observations.

In this paper, we first describe the technical problems caused by missing observations in performance. Then we present some common approaches used to overcome these missing data and the limitations of these approaches. Next, we discuss G-theory techniques, followed by an illustration of how to restructure and analyze a hypothetical sparsely-filled data set so that it can be accommodated by currently-available analytic methods. Finally, we apply our methods to a data set coming from a large scale writing assessment, and present the results of these analyses in terms of the comparability of the methods.

Theoretical Rationale

Because of a variety of problems unique to performance assessments (e.g., the extended amount of time required for examinees to formulate a response, the increased cost of testing, rater attrition, and rater availability), examinees may not respond to all items, and raters rarely evaluate all examinees. Brennan, Jarjoura, & Deaton (1980) refer to this situation as an unbalanced design, or a design with missing data. We adopt the latter term in this paper. Unfortunately, software that is designed to perform generalizability analyses, like GENOVA (Crick & Brennan, 1983) cannot handle missing data. Furthermore, according to Bell (1985) and Brennan (1992a), alternative analysis procedures (e.g., proc VARCOMP in SAS) that use iterative estimation methods (e.g., Maximum Likelihood or Restricted Maximum Likelihood) are computationally complex and require considerable computer resources and computational time. For example, Bell (1985) analyzed a survey containing the responses of 831 students from 112 schools with each student answering 11 questions (each response constituting a separate record so the total number of records were 11 x 831 = 9141). Bell compared the process time of two procedures for estimating variance components using the SAS system (SAS Institute, Inc., 1985), namely the VARCOMP procedure (i.e., TYPE1 and ML) and the GLM procedure. In all cases, a minimum of five minutes of central processing unit time was needed to complete the estimation procedure. We had a similar experience with the data analyzed for this study. At one point, we allowed the SAS VARCOMP procedure to run for over 24 hours on these data, and the estimation procedures still did not converge. In an age when funding for education is at a

premium, we must all find ways to conserve resources and avoid discarding data simply because we lack the technology to perform the analyses.

Researchers who use generalizability theory have devised several methods for analyzing test data with missing values. One approach is to collapse ratings across raters, ignoring the fact that different raters assigned scores to different examinees. For example, when two raters are randomly selected from a pool of raters to score examinees' response, it is a common practice to correlate the scores assigned by the first randomly-selected scorer with the scores assigned by the second randomly-selected scorer. The problem is that this approach jeopardizes the internal validity of the study by confounding the influences of multiple raters. A second approach is to select a single fully-crossed subset of data from the entire data set. An example of this approach may occur when a small number of raters make up a pool of raters from which pairs of raters are randomly assigned to score an examinee's response. In such a case, each pair of raters scores a small number of examinees in common. The pair of raters with the largest number of examinees in common may be chosen as the target of the analyses in such a situation. Unfortunately, by ignoring large portions of the data, this approach jeopardizes the external validity of the study (i.e., the chosen pair of raters may not be representative of the universe of raters). A third approach that may be employed is to perform analyses on all such fully-crossed subsets of data within a large data set and make comparisons across these data sets. Although this approach is considerably more desirable than the previous two, it still fails to take full advantage of all of the information contained in the entire data set. Our study investigates one option for analyzing missing data

that preserves both the internal and external validity of the G study while more fully utilizing the information contained in the data set.

Generalizability Theory

Generalizability theory offers a method of evaluating the effects that multiple sources of variability have on test reliability. Each source of variability is associated with a condition of the measurement framework called a facet (e.g., raters, items) or an interaction of these conditions (e.g., rater-by-item interactions). In this sense, G-theory extends the concept of measurement error as represented by classical test theory (i.e., Observed Score = True Score + Error) by decomposing the error term into multiple components that are associated with distinct features of the measurement context. In a two facet generalizability study, there are seven such terms. One facet arises from differences among examinees' performance and is denoted $\sigma^2(p)$. Typically, this facet is referred to as the object of measurement (Brennan, 1992a; Cronbach, Gleser, Nanda, & Rajaratnam; 1972, and Shavelson & Webb 1991). The second source of variability arises from the differences in the difficulty of the items, and is denoted as $\sigma^2(i)$. The third source of variability arises from the differences between the standards used by different raters. The fourth source of variability arises from the educational and experiential background that examinees bring to the test items. For instance, a test item could be more difficult for one student but not for the others. This examinee-by-item interaction is denoted $\sigma^2(pi)$. Two other sources of variability exist due to interactions between facets. The $\sigma^2(pr)$ represents the interaction due to the fact that different raters may apply the scoring criteria differentially across examinees (i.e., a rater-by-examinee interaction), whereas the $\sigma^2(ir)$ represents the interaction due to the fact that some raters apply

different standards to items that have the same level of difficulty (i.e., a rater-by-item interaction). The seventh source of variability may arise out of randomness, other systematic but unidentified error, or both. It is signified as σ^2(pir,e). This term is often referred to as the examinee-by-item-by-rater interaction, confounded by error.

One purpose of G-theory is to estimate the relative magnitude of indices (referred to as variance components) of the various sources of variability contained in a measurement context. This purpose is achieved through the use of a generalizability study (G study). Researchers examine the pattern and magnitude of these sources of variability and may change the scoring procedure with the hopes of reducing sources of error that are considered to be undesirable (e.g., the rater-related effects like σ^2(r), σ^2(pr), and σ^2(ir)) so that reliability can be increased. Measurement error attributable to effects like these can be reduced by increasing the number of raters, the number of items in a test, or both. A decision study (D study) is often used to estimate how changes in the number of items and/or number of raters would improve the reliability of an examinee's score. That is, D studies use the information from a G study concerning the multiple sources of measurement error to make projections to other operational settings. Introductory G-theory textbooks and research reports (e.g., Cronbach, Linn, Brennan, & Haertel, 1995; Shavelson & Webb, 1991) provide detailed discussions on the distinctions between these two studies. Our intent is to emphasize that getting valid information from a G study is critical to the precision of the projections that are made in a D study. The more comprehensive and representative the data we analyze, the more accurate and precise our predictions will be. However, the issue of representativeness is often treated as an assumption rather than as an empirical question. Our study examines this

representativeness issue by comparing the variance components obtained through different methods for compensating for missing data in a G study design.

Methods for Analyzing Missing Data

This section illustrates how to restructure a hypothetical sparsely-filled data matrix into smaller subsets for generalizability analyses. Our goal is to show how nearly all of the information can be considered without resorting to "discarding" data or ignoring distinctions by "collapsing" across elements (e.g., individual raters) of the measurement design. The G study results from the various designs that we describe can be averaged to produce a single set of variance components for the entire data set. In the following examples, we describe a measurement context in which 15 examinees each answer 2 test items which are rated by any 2 of 4 raters (named A, B, C, and D). That is, the design of our G study contains 2 facets: (a) items and (b) raters and can be represented as a fully-crossed examinee x item x rater (15 x 2 x 4) design with many pieces of missing data. Figure 1 depicts such a data matrix. This design matrix indicates which two raters rated a particular examinee on the two items. For instance, the four scores in the first row show that Examinee 1 was graded by Rater A and Rater B on both of the two items.

We can use four different methods to extract information from this sparse data matrix. In the collapsed method, we intentionally ignore which specific rater was the first rater or the second rater. That is, regardless of which pair of raters rated an examinee's response, the first rater in the pair was always labeled Rater 1 and the second rater was labeled Rater 2. Figure 2 depicts this collapsed data structure. The remaining three methods decompose the entire data set into exhaustive subsets. That is, the sum of the number of

observations analyzed by these three methods will equal the number of observations in the original data set. The data matrices from these subsets of data can each be analyzed under a different G study design. For the crossed method, we extract all possible crossed data subsets from the larger data set so that in each subset of data contains all examinees who were rated by a specific pair of raters on both items. In Figure 1, for example, Rater A and Rater B rated the response to both the first and the second item (AB, AB; where Item 1 and Item 2 are separated by a comma) for Examinees 1, 3, and 14. In this crossed design, each data set only contains scores given by a single pair of raters. Figure 3 shows how the information in Figure 1 decomposes into several crossed data sets. Responses of Examinees 1, 3, and 14 are rated by the same two raters (i.e., A & B) on both items, and for this reason scores for these two examinees are extracted from the entire data set and are stored in a smaller subset containing scores given by only Rater A and Rater B. In the same figure, Examinees 2 and 4 are graded by both Rater C and Rater D on the two items, and so these cases are extracted and saved in a data set with the label Crossed (2). The "2" in parentheses indicates the data set is a second of the crossed design type. In general, the parenthetical numbers distinguish one data set from the others within a type of design. By going down the rows, we exhaust all crossed designs and store them in these two subsets.

A nested design is formed every time one pair of raters rates the first item and a completely different rater pair rates the second item (e.g., Rater A and Rater B rate Item 1 and Rater C and Rater D rate Item 2 denoted AB, CD). Using the same algorithm as for extracting crossed data sets, we extract all nested subsets so that each nested data set contains scores of examinees' who are graded by the same four raters. In Figure 3, Examinees 5 and 6

are graded by Raters A and B on Item 1 and Raters C and D on Item 2. As a result, these examinees' scores are stored in one data set which is labeled Nested (1). Similarly, the Nested (2) data set contains all examinees scored by Raters B and C on Item 1 and Raters A and D on Item 2. The Nested (3) data set contains all examinees graded by Raters A and C on Item 1 and Raters B and D on Item 2. These three nested data sets exhaust all of the cases of nested ratings in the entire data set. Hence, we have used 12 of the 15 cases with the nested and crossed designs.

Our third design, the mixed design, accounts for the remaining cases. A mixed design is formed every time one rater rates both items and is paired with a different second rater on each item. For example, for Examinee 11, Rater A rates both items and is paired with Rater B on Item 1 and Rater C on Item 2 (AB, AC). However, there is a problem with this design--rater B and rater C always rate only one item each so that no information is available for evaluating the item effect for Rater B or Rater C. This problem is resolved by adding into the same data set two other rater combinations, (BA, BC) and (CA,CB). In these two designs rater B and rater C rate examinees' responses on both items. As a result, a fully nested data set contains all nested examples for a particular triplet of raters. Figure 3 depicts how to identify these three sets of raters and how to extract them from a data set. Because Examinees 11, 12, and 13 are, in turn, double-graded by the raters A, B, and C, the scores of these three examinees were stored into one data set and this data set also contains scores for other examinees who are graded by the same three raters in this mixed design.

Thus, we have been able to recover the data within the larger data set by parsing it into several subsets. Each of these subsets, if analyzed separately, will produce a set of variance component estimates. But, the variance component estimates produced by any one of these separate analyses may not adequately represent the variance structure of the entire data set. Unfortunately, the entire data set usually cannot be adequately analyzed because of weaknesses in the technology (i.e., computational time) or software (i.e., failure to handle missing data) used to perform the analyses. However, we can average the variance components from several G studies (Brennan, Gao, & Colton, 1995) to get more accurate and comprehensive variance component estimates. Hence, we can use our exhaustive parsing method (as described above) to extract all cases from a data set, perform G studies on each of these subsets of the data, and average the variance components across these G studies. These averaged variance components can serve as the information upon which D studies are based. In doing so, we preserve all of the information from the larger data set and create data sets with a structure that can be handled by currently-available software with a minimal processing load.

Research Questions

To our knowledge, such a method for parsing a large data set into mutually-exclusive and exhaustive subsets for the purpose of creating multiple data sets that are fully-analyzable in a generalizability theory framework has not been proposed. The goal of this approach is to obtain the most accurate G study variance component estimates possible so that generalizations beyond that data set will be valid. However, the validity of such a

method must be examined first. To this end, we investigated the following research questions.

1. Can these parsing methods be used to feasibly overcome the problem of computational complexity of analyzing a large, sparse data set?
2. Do the various methods produce comparable variance component estimates?
3. Are the variance components produced by the parsing methods superior to those produced by the collapsing method?

Method

The data we analyzed come from a large-scale college level writing assessment in which each examinee (N=5,905) responded to two essay prompts. Throughout the paper, we use "item" interchangeably with "essay". Each response was evaluated by a pair of raters randomly selected from a pool of nine trained raters, resulting in a total of 23,620 ratings (5,905 examinees x 2 essay prompts x 2 raters). Ratings were assigned on a six-point holistic scale. Table 1a summarizes the interrater agreement for the two essays. The total number of responses read by a particular rater ranged from 154 to 5,681 (see Table 1b for number of essays read by the nine raters). Because pairs of raters were randomly selected from a pool, this data set is sparse (i.e., not all examinees were rated by all raters). We analyzed the data using the four methods mentioned earlier.

All 23,620 ratings were analyzed using the collapsed design. A data set was analyzed as a crossed, nested, or mixed design only if the sample size for that data set was 20 or larger. As a result, we analyzed 9 of the 16 crossed data sets found in the entire data set, 22 of the 96 nested data sets, and 21 of the 40 mixed data sets. We used 84% (19,856) of the ratings, and

no data were used more than once. A G study was run for each of these data sets. GENOVA was used for the estimation of collapsed, crossed, and nested designs. The SAS VARCOMP procedure (estimation method = MIVQUE0) was used to analyze each mixed data set. According to Bell (1985), the MIVQUE0 estimation method is preferred to the TYPE1, ML, and REML because it is computationally efficient and the estimates are virtually identical to those obtained from full data sets. Because the sample sizes (i.e., number of examinees) vary among data sets, the variance components of the three designs were averaged using a pooled average formula modified from the formula typically used for obtaining the average over two samples. The formula follows.

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2}{(n_1 - 1) + (n_2 - 1) + \ldots + (n_k - 1)} \qquad (1)$$

where s_i^2 is the variance component in the ith data set within a design, n_i is the number of examinees in the ith data set within a design. For instance, there were nine data sets of the crossed design and therefore, the average variance component $\sigma^2(r)$ associated with the rater facet was a pooled average of the nine variance components from these data sets.

We evaluated the comparability of the three methods using these averaged variance components, based on the assumption that these aggregate components are representative of the individual data sets. This assumption was checked using the Pearson product moment correlation coefficients. In each design, we obtain a correlation between the average variance components with every single data set. We then took the average of these correlations. As a result, every one of the three designs has an average correlation which

indicates the extent to which the average variance components are representative of the individual data sets.

Following the assumption checking, we used a multivariate analysis of variance test (MANOVA) to examine the means of the variance components across the three methods. Wilk's λ is the test statistic associated with the MANOVA test and its distribution could be approximated by using Rao's F (Stevens, 1996). In our multivariate analysis, we hypothesize that the three methods are comparable in terms of the averaged variance components, and our null hypothesis in the multivariate analysis was that variance components were equal across the three methods. In the multivariate analysis, each variance component is treated as a dependent variable and the three methods are treated as levels in a factor. Although our intention is to conduct an omnibus test for the averaged variance components, the fact that the nested data sets have fewer variance components (due to the confounding rater and item effect) than the other two methods prohibits the use of a single multivariate test. To resolve this problem, we use two multivariate tests, one for comparing the seven variance components (see Table 8 for the seven components) between the crossed design and the mixed design, and the one for comparing the five variance components (see Table 9 for the five components) among all three designs.

To compare the crossed, mixed, and nested methods with the collapsed method, we conduct multiple one sample independent t-tests and adjust for the alpha level using the Bonferroni (e.g. Stevens, 1996) approach. Since we use 23 t-tests, we adjust the alpha level to .0021 (which is obtained by dividing the conventional level .05 by 23). Shavelson (1988) refers to the independent t-test we use a case 1 t-test. It is defined as

$$t_{observed} = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{N}}} \quad (2)$$

where $t_{observed}$ has degrees of freedom N-1, $\overline{X}$ is the average variance component within a facet, μ is the fixed value from the collapsed design, s and N are the standard deviation of the average variance component and the number of data sets used in a design, respectively. Using the t-test, each averaged variance component from the three designs was compared to the corresponding variance component obtained in the collapsed design. These independent t-tests indicate whether the sample mean of the variance component was drawn from a hypothesized population with a specified mean equal to the fixed value obtained from the collapsed design.

Results

Variability Within Parsing Methods

Table 2 shows the variance components for three of the crossed data sets chosen to be representative of the range of results obtained from the nine crossed data sets. In each case the $\sigma^2(p)$, $\sigma^2(pi)$, and $\sigma^2(pir,e)$ effects account for the greatest proportion of variance. The $\sigma^2(i)$ variance components are not as large, and the $\sigma^2(r)$, $\sigma^2(pr)$ and $\sigma^2(ir)$ components are negligible. However, there is considerable variability among these three crossed data sets. For example, the proportion of variance accounted by the $\sigma^2(p)$ effect ranged from 23% to 62% of the total variance. Such variability between subsets of the data emphasizes the risk associated

with estimating variance components from only a sample of raters. One way to avoid obtaining non-representative variance component estimates is to average variance components from multiple G studies (Brennan, Gao, & Colton, 95). Table 3 shows the variance components averaged across all nine of the crossed data sets we analyzed. The relative magnitudes of the variance components associated with each effect are similar to those in the individual data sets. However, these averaged variance components are more accurate and stable estimates of the variance components for the entire data set. Note that these averaged variance components have a rank ordering similar to that observed for each of the three example data sets shown in Table 2. The average correlation between these averaged variance components and the variance components obtained from each of the nine crossed data sets we analyzed was $\bar{r} = .91$.

Table 4 shows the estimated variance components for the three of the 21 mixed data sets. These were chosen to represent the range of results obtained under this parsing method. As with the crossed data sets, the largest variance are $\sigma^2(p)$, $\sigma^2(pi)$, and $\sigma^2(pir,e)$. The $\sigma^2(i)$ and $\sigma^2(pr)$ components are small. The $\sigma^2(r)$ and $\sigma^2(ir)$ terms are close to zero. There is a large amount of variability between the variance components obtained from the three example data sets for the mixed designs (as was true for the crossed data sets). Table 5 shows the average variance components across all 21 mixed data sets. Again, these estimates should be more representative of the information contained in the entire data set than would be any single subset of the data. The averaged variance components shown in Table 5 are similar to those obtained from each of the individual mixed data sets. The average correlation was $\bar{r} = .94$.

Table 6 shows the variance components for three of the 22 nested data sets. Comparing to the other two designs, the nested design has more data sets of smaller size. Recall that only 22 of the 96 data sets have 20 or more examinees. This is not surprising in operational settings in which it is more convenient and efficient to randomly select four different raters than to systematically pair up raters. Due to the fact that raters are confounded within essays, there is no way of estimating the unique effect for $\sigma^2(r)$, $\sigma^2(ir)$, and $\sigma^2(pr)$. This confounding nature allows estimations to be made for only five variance components. Like the other two designs, the largest proportion of variance is accounted by $\sigma^2(p)$, $\sigma^2(pi)$, and $\sigma^2(p(r:i),e)$. The within variability among the data sets is large that, for example, the $\sigma^2(p)$ is ranged from 11.33% to 65.71%. Again, the average variance components are more representative to the entire data set. The average correlation between these averaged variance components and the variance components obtained from each of the 22 nested data sets we analyzed was .76. Although this $\bar{r}$ seems low comparing to those obtained for the crossed and mixed designs, when considering the fact that the nested design has two variance components fewer than the other two designs, a correlation of .76 suggests that the average variance components resemble a reasonably consistent rank ordering of the individual data sets.

Comparison Across Parsing Methods

Table 8 compares the average variance components for the crossed and mixed data sets with the variance components obtained from the collapsed method. The proportions of variance accounted for by each effect in these three designs are similar. As would be expected, the $\sigma^2(r)$, $\sigma^2(pr)$, and $\sigma^2(ir)$ effects look smaller with the collapsed method, as the result of confounding the effects of individual raters. Independent t-tests indicate that there

are no statistically significant differences between the variance components from the mixed and crossed designs and those from the collapsed design. A MANOVA test is conducted to test if any pairs of the mean variance components differ between the crossed and mixed methods. The omnibus test is insignificant (Wilk's $\lambda = .79$, $F_{7,22} = 0.83$, $p = .57$) implying that no mean variance components differ between the two designs.

Table 9 compares variance components from the crossed, mixed, and nested methods to those from the collapsed method. Because the $\sigma^2(r)$ and $\sigma^2(i)$ effects are confounded in the nested design, it is necessary to recalculate the variance components for the previous designs to show such a similar confounding effect. To this end, the $\sigma^2(r:i)$ component is estimated as the sum of the $\sigma^2(r)$ component and the $\sigma^2(ir)$ components, and the $\sigma^2(p(r:i),e)$ component is estimated as the sum of the $\sigma^2(pr)$ and $\sigma^2(pir,e)$ components (Brennan, 1992b) for the average variance components obtained under the crossed, mixed, and collapsed methods. For the nested method, the variance components for $\sigma^2(p)$ and $\sigma^2(p(r:i),e)$ are slightly smaller, and the $\sigma^2(i)$ and $\sigma^2(pi)$ variance components are slightly larger than those of the crossed, mixed, and collapsed designs. An omnibus MANOVA test reveals that none of these differences, in the nested method, are large enough to be considered significant (Wilk's $\lambda = .77$, $F_{10,90} = 1.27$, $p = .26$). However, using the one sample independent t-tests, we find $\sigma^2(r:i)$ differs significantly ($p < .002$) between the collapsed design and the mixed design. We also find that $\sigma^2(p(r:i),e)$ differs significantly ($p < .002$) between the collapsed and the nested design.

Discussion and Conclusions

In this paper, we have shown that sparsely-filled performance assessment data sets can be restructured into analyzable smaller subsets of data. The method we used is particularly suitable for analyzing operational performance assessment data in which missing observations are unavailable due to the constraints caused by using expert judgments for scoring or by the increased costs of administering these assessments.

As opposed to our expectation, the results obtained from the collapsed method look similar to that from the other three methods. However, we recommend against the use of this collapsed method because we know it is incorrect to ignore raters' identity. One possible explanation to these unanticipated results is that the t-tests we used are inappropriate for comparing the collapsed method to the other three methods. It is inappropriate because the data analyzed in the three methods are dependent on those analyzed in the collapsed method (i.e. the examinees in each of the three methods are the same as those analyzed in the collapsed method). A more appropriate test is needed in future studies for handling this dependency issue.

Although our results indicate that the three methods (i.e., crossed, mixed, and nested) are comparable using a large scale writing assessment data set, we need to conduct a more thorough study to examine whether the methods generalize to other data sets with more than two facets and to other data sets with different magnitude in the variance components. Based on the evidence, the averaged variance components across designs are probably the best estimate of "true" variance components for a sparsely-filled data set. Future Monte Carlo (MC) studies (e.g., Hamilton, 1992, provides a lucid introduction to

this topic; Harwell, 1992, discusses how to summarize MC results in methodological research) are need to be used to determine whether each method produces unbiased and accurate variance component estimates. This could be accomplished by generating a large, fully-crossed data set, following by obtaining the variance components from computer packages such as GENOVA. Then, we could randomly pull samples from this large data set according to the specifications of our rating design (i.e., we would randomly-sample 2 raters for every examinee x item combination). For each of these samples, we would estimate variance components using the various methods for parsing the data. This would create a distribution of variance component estimates for each facet of the design under each of our parsing methods. We would check if the different methods are unbiased by examining how different the means of the distributions differs from the parameter value, and we would know if the different methods are accurate by examining the variance of the distributions around the parameter value.

Last but not least, for future studies, we suggest researchers explore the use of other tests, in addition to the MANOVA test, for examining the comparability of the three methods we employed. If the methods are comparable, no matter what tests are used the results should be the same – that variance components from the three methods do not differ significantly within a facet. One such test for comparability is the multivariate homogeneity test (Raudenbush, Becker and Kalaian, 1988) which is usually used in research synthesis (or meta-analysis coined by Glass, 1976). The advantage of this multivariate homogeneity test is that only one analysis is needed for testing whether or not the averaged variance components are representative and comparable.

Once the three methods are tested to be comparable using a variety of tests (e.g. MANOVA test and homogeneity test) AND once they are shown to be generalizable to other data sets with more than two facets, we could consider taking a step further to average the mean variance components obtained from the three methods. This way of taking an average could be applied to every variance component and, therefore every component has an index. We expect these average mean variance components to be the most stable and parsimonious indices in generalizability studies in which some observations are missing.

References

Bell, J. F. (1985). Generalizability theory: The software problem. Journal of Educational Statistics, 10(1), 19-29.

Brennan, R. L. (1992a). Elements of Generalizability Theory. Iowa City, Iowa: American College Testing.

Brennan, R. L. (1992b). An NCME instructional module on Generalizability theory. Educational Measurement: Issues and Practice, 11(4), 27-34.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of work keys listening and writing tests. Educational and Psychological Measurement, 55(2), 157-176.

Brennan, R. L., Jarjoura, D., & Deaton, E. L. (1980). Some issues concerning the estimation and interpretation of variance components in Generalizability Theory. ACT Technical Bulletin. (Report No. Number 36). Iowa City, Iowa: American College Testing.

Crick, J. E., & Brennan, R. L. (1982). GENOVA: A generalized analysis of variance system (FORTRAN IV computer program and manual). Dorchester, Mass: Computer Facilities, University of Massachusetts at Boston.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons, Inc.

Cronbach, L. J., Gleser, G. C., & Rajaratnam, N. (1963). Theory of generalizability. A liberalization of reliability theory. British Journal of Mathematical and Statistical Psychology, 16, 137-173.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel E. (1995). Generalizability analysis for educational assessments. Evaluation Comment. Los Angles, Center for the Study of Evaluation, UCLA.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5(10), 3-8.

Hamilton, L. C. (1992). Regression with Graphics. Computer-intensive methods (Appendix 2) . Belmont, California: Duxbury Press.

Harwell, M. R. (1992). Summarizing Monte Carlo results in methodological research. Journal of Educational Statistics, 17(4), 297-313.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. Psychological Bulletin, 103(1), 111-120.

SAS Institute, Inc. (1985). SAS User's Guide: Statistics. (Version 5 ed.). Cary, NC: Author.

Shavelson, J. R. (1988). Statistical Reasoning for the Behavioral Sciences. (2nd ed.). Boston: Allyn and Bacon.

Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage.

Stevens, J. (1996). Applied multivariate statistics for the social sciences. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Figure 1. A hypothetical data matrix. Rows represent examinees, columns represent items, and sub-columns represent raters. This matrix contains scores assigned by 2 randomly-selected raters (from the pool of four) to each examinee's response to each item. Subscripts represent examinees, items, and raters.

Figure 2. A hypothetical data matrix that illustrates how the data are structured when rater identity is ignored. Subscripts represent examinees, items, "collapsed" rater, and "specific" rater. In a collapsed analysis, the "specific" rater is ignored.

Figure 3. A hypothetical data matrix, identical to that in Figure 1. The last two columns illustrate how to restructure the sparse data set into multiple smaller data sets analyzable by the three methods, namely the crossed, mixed, and nested methods.

Figure 1

		Item 1				Item 2			
		Rater A	Rater B	Rater C	Rater D	Rater A	Rater B	Rater C	Rater D
Examinee	1	$X_{1,1,A}$	$X_{1,1,B}$			$X_{1,2,A}$	$X_{1,2,B}$		
	2			$X_{2,1,C}$	$X_{2,1,D}$			$X_{2,2,C}$	$X_{2,2,D}$
	3	$X_{3,1,A}$	$X_{3,1,B}$			$X_{3,2,A}$	$X_{3,2,B}$		
	4			$X_{4,1,C}$	$X_{4,1,D}$			$X_{4,2,C}$	$X_{4,2,D}$
	5	$X_{5,1,A}$	$X_{5,1,B}$					$X_{5,2,C}$	$X_{5,2,D}$
	6	$X_{6,1,A}$	$X_{6,1,B}$					$X_{6,2,C}$	$X_{6,2,D}$
	7		$X_{7,1,B}$	$X_{7,1,C}$		$X_{7,2,A}$			$X_{7,2,D}$
	8	$X_{8,1,A}$		$X_{8,1,C}$			$X_{8,2,B}$		$X_{8,2,D}$
	9	$X_{9,1,A}$		$X_{9,1,C}$			$X_{9,2,B}$		$X_{9,2,D}$
	10		$X_{10,1,B}$	$X_{10,1,C}$		$X_{10,2,A}$			$X_{10,2,D}$
	11	$X_{11,1,A}$	$X_{11,1,B}$			$X_{11,2,A}$		$X_{11,2,C}$	
	12	$X_{12,1,A}$	$X_{12,1,B}$				$X_{12,2,B}$	$X_{12,2,C}$	
	13	$X_{13,1,A}$		$X_{13,1,C}$			$X_{13,2,B}$	$X_{13,2,C}$	
	14	$X_{14,1,A}$	$X_{14,1,B}$			$X_{14,2,A}$	$X_{14,2,B}$		
	15		$X_{15,1,B}$	$X_{15,1,C}$		$X_{15,2,A}$			$X_{15,2,D}$

BEST COPY AVAILABLE

Figure 2

		Item 1		Item 2	
		Rater 1	Rater 2	Rater 1	Rater 2
Examinee	1	$X_{1,1,1,A}$	$X_{1,1,2,B}$	$X_{1,2,1,A}$	$X_{1,2,2,B}$
	2	$X_{2,1,1,C}$	$X_{2,1,2,D}$	$X_{2,2,1,C}$	$X_{2,2,2,D}$
	3	$X_{3,1,1,A}$	$X_{3,1,2,B}$	$X_{3,2,1,A}$	$X_{3,2,2,B}$
	4	$X_{4,1,1,C}$	$X_{4,1,2,D}$	$X_{4,2,1,C}$	$X_{4,2,2,D}$
	5	$X_{5,1,1,A}$	$X_{5,1,2,B}$	$X_{5,2,1,C}$	$X_{5,2,2,D}$
	6	$X_{6,1,1,A}$	$X_{6,1,2,B}$	$X_{6,2,1,C}$	$X_{6,2,2,D}$
	7	$X_{7,1,1,B}$	$X_{7,1,2,C}$	$X_{7,2,1,A}$	$X_{7,2,2,D}$
	8	$X_{8,1,1,A}$	$X_{8,1,2,C}$	$X_{8,2,1,B}$	$X_{8,2,2,D}$
	9	$X_{9,1,1,A}$	$X_{9,1,2,C}$	$X_{9,2,1,B}$	$X_{9,2,2,D}$
	10	$X_{10,1,1,B}$	$X_{10,1,2,C}$	$X_{10,2,1,A}$	$X_{10,2,2,D}$
	11	$X_{11,1,1,A}$	$X_{11,1,2,B}$	$X_{11,2,1,A}$	$X_{11,2,2,C}$
	12	$X_{12,1,1,A}$	$X_{12,1,2,B}$	$X_{12,2,1,B}$	$X_{12,2,2,C}$
	13	$X_{13,1,1,A}$	$X_{13,1,2,C}$	$X_{13,2,1,C}$	$X_{13,2,2,B}$
	14	$X_{14,1,1,A}$	$X_{14,1,2,B}$	$X_{14,2,1,A}$	$X_{14,2,2,B}$
	15	$X_{15,1,1,B}$	$X_{15,1,2,C}$	$X_{15,2,1,A}$	$X_{15,2,2,D}$

BEST COPY AVAILABLE

Figure 3

		Item 1				Item 2				Data Set	Design (File ID)
		Rater A	Rater B	Rater C	Rater D	Rater A	Rater B	Rater C	Rater D		
Examinee	1	$X_{1,1,1}$	$X_{1,1,2}$			$X_{1,2,1}$	$X_{1,2,2}$			(AB, AB)	Crossed (1)
	2			$X_{2,1,1}$	$X_{2,1,2}$			$X_{2,2,1}$	$X_{2,2,2}$	(CD, CD)	Crossed (2)
	3	$X_{3,1,1}$	$X_{3,1,2}$			$X_{3,2,1}$	$X_{3,2,2}$			(AB, AB)	Crossed (1)
	4			$X_{4,1,1}$	$X_{4,1,2}$			$X_{4,2,1}$	$X_{4,2,2}$	(CD, CD)	Crossed (2)
	5	$X_{5,1,1}$	$X_{5,1,2}$					$X_{5,2,1}$	$X_{5,2,2}$	(AB, CD)	Nested (1)
	6	$X_{6,1,1}$	$X_{6,1,2}$					$X_{6,2,1}$	$X_{6,2,2}$	(AB, CD)	Nested (1)
	7		$X_{7,1,1}$	$X_{7,1,2}$		$X_{7,2,1}$			$X_{7,2,2}$	(BC, AD)	Nested (2)
	8	$X_{8,1,1}$		$X_{8,1,2}$			$X_{8,2,1}$		$X_{8,2,2}$	(AC, BD)	Nested (3)
	9	$X_{9,1,1}$		$X_{9,1,2}$			$X_{9,2,1}$		$X_{9,2,2}$	(AC, BD)	Nested (3)
	10		$X_{10,1,1}$	$X_{10,1,2}$		$X_{10,2,1}$			$X_{10,2,2}$	(BC, AD)	Nested (2)
	11	$X_{11,1,1}$	$X_{11,1,2}$			$X_{11,2,1}$		$X_{11,2,2}$		(AB, AC)	Mixed (1)
	12	$X_{12,1,1}$	$X_{12,1,2}$				$X_{12,2,1}$	$X_{12,2,2}$		(AB, BC)	Mixed (1)
	13	$X_{13,1,1}$		$X_{13,1,2}$			$X_{13,2,1}$	$X_{13,2,2}$		(AC, CB)	Mixed (1)
	14	$X_{14,1,1}$	$X_{14,1,2}$			$X_{14,2,1}$	$X_{14,2,2}$			(AB, AB)	Crossed (1)
	15		$X_{15,1,1}$	$X_{15,1,2}$		$X_{15,2,1}$			$X_{15,2,2}$	(BC, AD)	Nested (2)

BEST COPY AVAILABLE

Table 1a

Average Percentage of Agreement for Two Essays

	Perfect Agreement	Percent Adjacent (1 Scale Point)	Percent Non-Adjacent (2 or more Scale Point)
Essay 1	73.6	25.5	0.9
Essay 2	73.6	26.2	0.3

Table 1b

Number of Essays Read by the Nine Raters

Rater	Essay 1 (Frequency)	Essay 2 (Frequency)	Total (Frequency)	Total (Percentage)
1	992	836	1828	7.7
2	2797	2884	5681	24.1
3	485	316	801	3.4
4	2281	2509	4790	20.3
5	2011	2002	4013	17
6	2169	2474	4643	19.7
7	100	130	230	1.0
8	856	624	1480	6.3
9	119	35	154	0.7
Total Number of Essays Read by the Nine Raters			= 23,620	100

Table 2

Variance Components for Three Crossed Design Data Sets

	Data Set 1 (N=144)		Data Set 2 (N=179)		Data Set 3 (N=61)	
Source	VC	%	VC	%	VC	%
Person (p)	0.33399	62.13	0.12771	36.81	0.08593	22.56
Item (i)	0.00918	1.71	0.04309	12.42	0.02828	7.42
Rater (r)	0.00009	0.02	0.00025	0.07	0.00403	1.06
Person x Item (pi)	0.08076	15.02	0.07981	23.00	0.16025	42.07
Person x Rater (pr)	0.00000	0.00	0.00534	1.54	0.00000	0.00
Item x Rater (ir)	0.00000	0.00	0.00176	0.51	0.00984	2.58
Person x Item x Rater, Error (pir,e)	0.11353	21.12	0.08902	25.66	0.09262	24.31

Note: N is the number of persons contained in the data set. VC is the variance component for the source. % is the percent of the total variance accounted for by the source.

Table 3

Average Variance Components for All Crossed Data Sets

Source	Mean VC	%	SD VC
Person (p)	0.21123	44.86	0.07989
Item (i)	0.02396	5.09	0.01851
Rater (r)	0.00186	0.39	0.00185
Person x Item (pi)	0.11199	23.78	0.03783
Person x Rater (pr)	0.01135	2.41	0.04185
Item x Rater (ir)	0.00222	0.47	0.00528
Person x Item x Rater, Error (pir,e)	0.10828	22.99	0.02277

Note: Mean VC is the average variance component across the nine crossed data sets. % is the percent of the total variance accounted for by the source. SD VC is the standard deviation of the variance component across the nine crossed designs. The average N=105.

Table 4

Variance Components for Three Mixed Design Data Sets

	Data Set 1 (N=158)		Data Set 2 (N=132)		Data Set 3 (N=52)	
Source	VC	%	VC	%	VC	%
Person (p)	0.27380	53.87	0.27597	43.03	0.07186	21.78
Item (i)	0.02062	4.06	0.03090	4.82	0.01619	4.91
Rater (r)	0.00003	0.01	0.00200	0.31	0.00000	0.00
Person x Item (pi)	0.11412	22.46	0.19736	30.77	0.13550	41.08
Person x Rater (pr)	0.00099	0.19	0.03986	6.21	0.01598	4.84
Item x Rater (ir)	0.00240	0.47	0.00000	0.00	0.00338	1.02
Person x Item x Rater, Error (pir,e)	0.09627	18.94	0.09527	14.85	0.08696	26.36

Note: N is the number of persons contained in the data set. VC is the variance component for the source. % is the percent of the total variance accounted for by the source.

Table 5

Average Variance Components for All Mixed Data Sets

Source	Mean VC	%	SD VC
Person (p)	0.20771	44.75	0.08296
Item (i)	0.01833	3.95	0.01857
Rater (r)	0.00268	0.58	0.00518
Person x Item (pi)	0.11745	25.30	0.04163
Person x Rater (pr)	0.01223	2.63	0.02111
Item x Rater (ir)	0.00287	0.62	0.01174
Person x Item x Rater, Error (pir,e)	0.10290	22.17	0.04794

Note: Mean VC is the average variance component across the 21 mixed data sets. % is the percent of the total variance accounted for by the source. SD VC is the standard deviation of the variance component across the 21 mixed designs. The average N=161.

Table 6

Variance Components for Three Nested Design Data Sets

	Data Set 1 (N=20)		Data Set 2 (N=21)		Data Set 3 (N=30)	
Source	VC	%	VC	%	VC	%
Person (p)	0.36053	65.71	0.20417	40.26	0.03621	11.33
Item (i)	0.03224	5.88	0.00000	0.00	0.01466	4.59
Rater:Item (r:i)	0.00000	0.00	0.01071	2.11	0.02069	6.47
Person x Item (pi)	0.06711	12.23	0.21964	43.31	0.12701	39.75
Person x (Rater:Item), Error (p(r:i))	0.08882	16.19	0.07262	14.32	0.12098	37.86

Note: N is the number of persons contained in the data set. VC is the variance component for the source. % is the percent of the total variance accounted for by the source.

Table 7

Average Variance Components for All Nested Data Sets

Source	Mean VC	%	SD VC
Person (p)	0.19170	41.62	0.09555
Item (i)	0.03613	7.84	0.05946
Rater:Item (r:i)	0.00314	0.68	0.00575
Person x Item (pi)	0.12995	28.21	0.07967
Person x Rater:Item, Error (p(r:i),e)	0.09967	21.64	0.02041

Note: Mean VC is the average variance component across the 22 nested data sets. % is the percent of the total variance accounted for by the source. SD VC is the standard deviation of the variance component across the 22 nested designs. The average N=29.

Table 8

Averaged Components for Crossed, Mixed, and Collapsed Parsing Methods

	Crossed		Mixed		Collapsed	
Source	Mean VC	%	Mean VC	%	Mean VC	%
p	0.21123	44.86	0.20771	44.75	0.21031	44.25
i	0.02396	5.09	0.01833	3.95	0.01784	3.75
r	0.00186	0.39	0.00268	0.58	0.00036	0.08
pi	0.11199	23.78	0.11745	25.30	0.12710	26.74
pr	0.01135	2.41	0.01223	2.63	0.00336	0.71
ir	0.00222	0.47	0.00287	0.62	0.00000	0.00
pir,e	0.10828	22.99	0.10290	22.17	0.11629	24.47

Note: Mean VC is the averaged variance component for the source across all data sets of this type. % is the percent of the total variance accounted for by the source. Associated statistics for the three rater-related variance components are as follows. In the crossed design: for $\sigma^2(r)$, $t_8 = 1.93$, $p = .09$; for $\sigma^2(pr)$, $t_8 = 1.25$, $p = .25$; for $\sigma^2(ir)$, $t_8 = 1.62$, $p = .15$. In the mixed design: for $\sigma^2(r)$, $t_{20} = 3.19$, $p = .005$; for $\sigma^2(pr)$, $t_{20} = 2.65$, $p = .016$; for $\sigma^2(ir)$, $t_{20} = 2.12$, $p = .05$.

Table 9

Averaged Components for Crossed, Mixed, Nested, and Collapsed Parsing Methods

	Crossed		Mixed		Nested		Collapsed	
Source	Mean VC	%	Mean VC	%	Mean VC	%	Mean VC	%
p	0.21123	44.86	0.20771	44.75	0.19170	41.62	0.21031	44.25
i	0.02396	5.09	0.01833	3.95	0.03613	7.84	0.01784	3.75
r:i	0.00408	0.87	0.00841*	1.20	0.00314	0.68	0.00036	0.08
pi	0.11199	23.78	0.11745	25.30	0.12995	28.21	0.12710	26.74
p(r:i),e	0.11963	25.40	0.11513	24.80	0.09967*	21.64	0.11965	25.18

Note: Mean VC is the averaged variance component for the source across all data sets of this type. % is the percent of the total variance accounted for by the source. Associated statistics for the two rater-related variance components are as follows. In the crossed design: for σ^2(r:i), $t_8 = 2.16$, $p = .06$; for σ^2((p(r:i)), $t_8 = 0.76$, $p = .47$. In the mixed design: for σ^2(r:i), $t_{20} = 3.76$, $p = .001$; for σ^2((p(r:i)), $t_{20} = 0.70$, $p = .49$. In the nested design: σ^2(r:i), $t_{20} = 2.30$, $p = .03$; for σ^2((p(r:i)), $t_{20} = 4.11$, $p < .001$.